USGS
science for a changing world

SEM Essentials: Estimation

Jim Grace

U.S. Department of the Interior
U.S. Geological Survey

In this module, I give a brief overview of the estimation methods used in SEM.

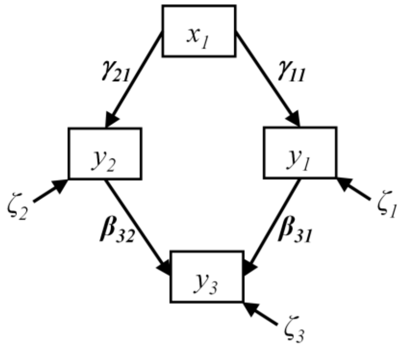
A citation that can be used for the information included in this module is

Grace, J.B., Schoolmaster, D.R. Jr., Guntenspergen, G.R., Little, A.M., Mitchell, B.R., Miller, K.M., and Schweiger, E.W. 2012. Guidelines for a graph-theoretic implementation of structural equation modeling. Ecosphere 3(8): article 73 (44 pages). http://www.esajournals.org/doi/pdf/10.1890/ES12-00048.1.

Notes: IP-056512; Support provided by the USGS Climate & Land Use R&D and Ecosystems Programs. I would like to acknowledge formal review of this material by Jesse Miller and Phil Hahn, University of Wisconsin. Many helpful informal comments have contributed to the final version of this presentation. The use of trade names is for descriptive purposes only and does not imply endorsement by the U.S. Government. Last revised 20141216. Questions about this material can be sent to sem@usgs.gov.

1. We can look at the estimation problem in a general way.

Consider a simple model where x_1 affects y_3 through two routes (via y_1 and y_2.



This model can be represented by three equations, one for each endogenous variable.

$$y_1 = \alpha_1 + \gamma_{11}x_1 + \zeta_1$$

$$y_2 = \alpha_2 + \gamma_{21}x_1 + \zeta_2$$

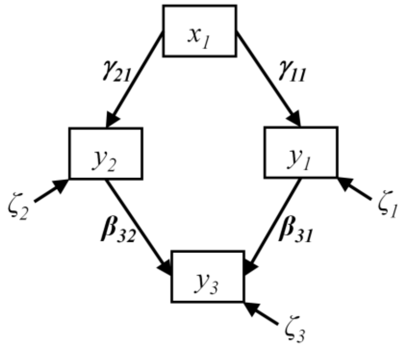
$$y_3 = \alpha_3 + \beta_{31}y_1 + \beta_{32}y_2 + \zeta_3$$

USGS

2

There are certain general features of estimation in SE models. Typically there are equations for each endogenous variable and parameters, such as intercepts and slopes, which we must estimate for the prediction equations, based on some method for maximizing explanation of the variations in the data relative to the predicted relationship. As part of the estimation process we derive additional estimates for parameter standard errors (or posterior distribution characteristics in Bayesian estimation), as well as error variances, residual covariances and more.

2. A key test to be performed that involves the parameter estimates is the test of <u>conditional independence</u>.



The model architecture implies that,
$COV(x_1, y_3) = \gamma_{11} * \beta_{31} + \gamma_{21} * \beta_{32}$.

If this is not true, there is some additional process connecting x_1 and y_3.

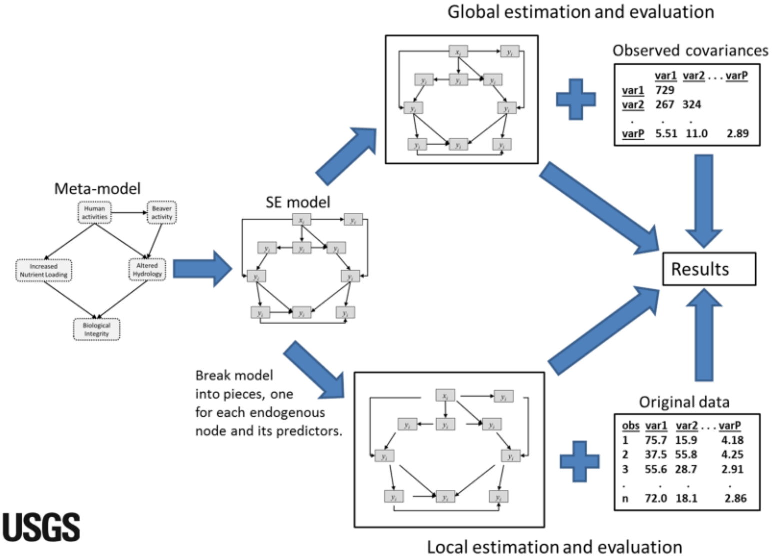
USGS

3

In addition to estimating model parameters, we also estimate in some fashion how well the data agree with the model. Ultimately, the model property being tested is whether the omitted links that are part of the hypothesis are consistent with the data relations. I go into this in more detail in the module on path rules, but basically, the question is whether estimated indirect effects in the model match with observed covariances. If not, that implies something is wrong with the model, such as a missing path directly from x_1 to y_3 in this case.

Note that there is one other implied conditional independence in the model besides the one between x_1 and y_3.

3. There are two basic approaches to estimating the parameters in a model. One is via local “piecewise” estimation, the other via global estimation.



USGS

4

This slide compares global versus local estimation. First, based on theoretical ideas and objectives, a meta-model is derived that represents general conceptual expectations for the ecological situation. From that, either directly or from a causal diagram, a structural equation model is developed for analysis. SE models can be analyzed either under a global framework or through piecewise estimation of local relationships. While analytical procedures differ, both approaches represent implementations of the SEM paradigm.

4. Local estimation involves estimating parameter values for each equation separately, then assembling the model as a collection of prediction equations.

$$y_1 = \alpha_1 + \gamma_{11} x_1 + \zeta_1$$

$$y_2 = \alpha_2 + \gamma_{21} x_1 + \zeta_2$$

$$y_3 = \alpha_3 + \beta_{31} y_1 + \beta_{32} y_2 + \zeta_3$$

USGS

5

Local estimation was the original approach in SEM involving path analysis and is making a resurgence (even advocated by Pearl for the general implementation of SEM). We are not all the way there yet with local estimation. There are some models we can’t yet estimate in this fashion, as described in the module on Model Specifications. Also, local estimation implementations are not yet automated, though we are working on that.

5. Second approach to estimating the parameters is through "global" estimation methods.

- Conceptualize model as collection of vectors (of variables, intercepts, and errors) and matrices (regression parameters, error correlations).

$$y_1 = \alpha_1 + \gamma_{11}x_1 + \zeta_1$$
$$y_2 = \alpha_2 + \gamma_{21}x_1 + 0*y_1 + \zeta_2$$
$$y_3 = \alpha_3 + 0*x_1 + \beta_{31}y_1 + \beta_{32}y_2 + \zeta_3$$

$$\Downarrow$$

$$\boxed{Y = \alpha + \Gamma X + \beta Y + \zeta}$$

Y = p x 1 vector of responses

α = p x 1 vector of intercepts

Β = p x p coefficient matrix of ys on ys

Γ = p x q coefficient matrix of ys on xs

X = q x 1 vector of exogenous predictors

ζ = p x 1 vector of errors for the elements of y

Φ = cov (X) = q x q matrix of covariances among xs

Ψ = cov (ζ) = q x q matrix of covariances among errors

USGS

6

Modern SEM is strongly associated with the global estimation paradigm. This was first developed by Kark Joreskog in the early 1970s and implemented in his LISREL software, which relied heavily on matrix algebra procedures.

6. Global estimation uses matrix methods.

Pre-analysis step: Summarize raw data in variance-covariance matrices.

data

Row	x	y_1	y_2	y_3
1	40	3.5	1.04	51
2	25	4.0	0.48	31
3	15	2.6	0.95	71
4	23	4.3	1.19	64
5	24	4.0	1.30	68
.	.	.	.	.
n	15	3.8	0.69	40

variance/covariance matrix*

	x	y_1	y_2	y_3
x	1			
y_1	-0.35	1		
y_2	0.45	-0.44	1	
y_3	-0.30	0.33	-0.37	1
std dev	12.6	0.32	1.65	15.1
mean	25.6	0.69	4.56	49.2

*showing standardized matrix

plus other summary information

USGS

7

A key feature of the global estimation approach is the summary of data relations in the form of a variance/covariance matrix. This is the “trick” that allowed factor-type models and econometric models with reciprocal causation to be estimated.

Other summary information that can be derived from the data includes things like kurtosis of variables, which can be used in adjustments employed in “robust” estimation.

7. The basic problem in global estimation is to estimate parameters by comparing observed covariances to model-implied covariances.

compare

Observed Covariances

$$S = \left\{ \begin{matrix} 159 & & & \\ -1.4 & 0.10 & & \\ 9.36 & 0.23 & 2.72 & \\ -57.1 & 0.11 & -0.93 & 228 \end{matrix} \right\}$$

Model-Implied Covariances

$$\Sigma = \left\{ \begin{matrix} \sigma_{11} & & & \\ \sigma_{12} & \sigma_{22} & & \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \sigma_{13} \end{matrix} \right\}$$

USGS

8

Essentially, by representing the problem in terms of comparing model-implied covariances with observed covariances, and using maximum likelihood techniques, it is possible to estimate the parameters for a wide variety of models, including those with latent variables and causal loops.

Parameter estimates are obtained by iterative comparisons.

8. There is a "fundamental equation" for global estimation.

The fundamental hypothesis behind covariance-based SEM is

$$\Sigma = \Sigma(\Theta)$$

where:
Σ = population covariance matrix of observed variables,
Θ = vector of population parameter values for the model, and
$\Sigma(\Theta)$ = the covariance matrix written as a function of Θ

In practice, we are dealing with estimates.

$$\hat{\Sigma} = \Sigma(\hat{\Theta})$$

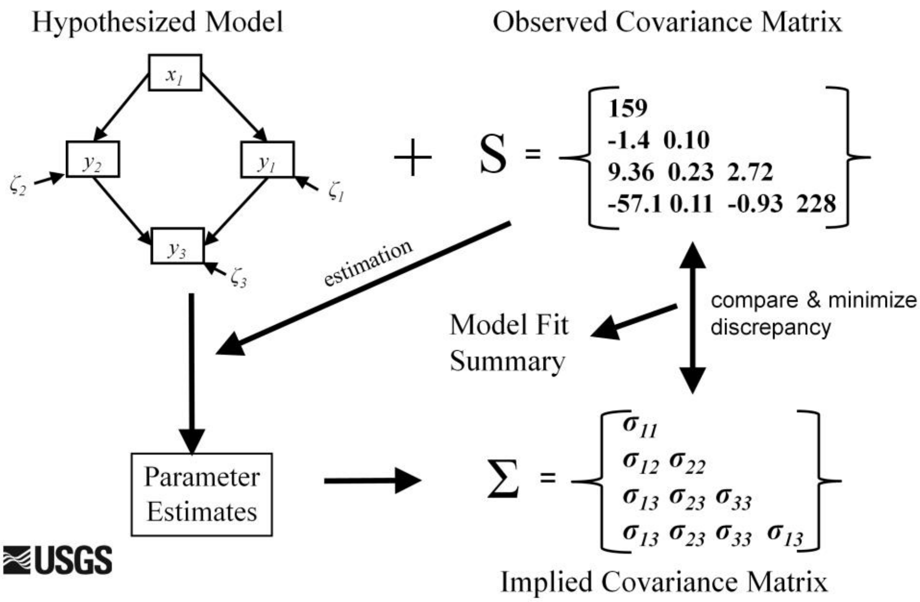
we aspire to $\hat{\Theta}$ such that $S = \hat{\Sigma}$

USGS

9

Our objective in global estimation is to pick values of parameter estimates such that the model-implied covariance matrix is as close to S, the sample covariance matrix, as possible. This just says that we try to make things add up.

9. We can think of global estimation as a matching process.



Here is a cartoon to represent the sequence of steps in global estimation.

10. The core of global estimation is the fitting function.

Fitting functions are designed to minimize model-data discrepancies.

Most common fitting function is based on the log likelihood ratio, which compares the likelihood for a given model to the likelihood of a model with perfect fit.

$$F_{ML} = \log\left|\hat{\mathbf{\Sigma}}\right| + tr\left(\mathbf{S}\hat{\mathbf{\Sigma}}^{-1}\right) - \log\left|\mathbf{S}\right| - (p+q)$$

Note that when sample matrix and implied matrix are equal, terms 1 and 3 = 0 and terms 2 and 4 = 0. Thus,

perfect model fit yields a value of F_{ML} of 0.

USGS

11

At the core of the computations are those that estimate the degree of discrepancy, overall, between the observed and model-implied covariances. This discrepancy estimate F_{ML} is an essential quantity in other calculations, as we shall see.

11. Maximum likelihood estimators, such as F_{ML}, possess several important properties:

(1) asymptotically unbiased,
(2) scale invariant, and
(3) best estimators.

Assumptions:

(1) Σ-hat and S matrices are **positive definite** (i.e., that they do not have a singular determinant such as might arise from a negative variance estimate, an implied correlation greater than 1.0, or from one row of a matrix being a linear function of another), and

(2) data follow a multinormal distribution.

USGS 12

At this level in SEM, we encounter statistical assumptions that accompany the estimation procedures used. In addition to the usual assumptions that go along with maximum likelihood, there is the requirement that the matrices are "positive definite". It is not uncommon in SEM analyses to receive an error message that there are "non-positive definite" elements in the matrices. This statement of a computational error implies usually some computable relations among the variables, such as one variable can be computed from a combination of the others or two variables very highly correlated (greater than 0.99).

12. Parameter "identification" is a key, fundamental topic.

1. For model parameters to be estimated with unique values, they must be identified. As in linear algebra, we have a requirement that we need as many known pieces of information as we do unknown parameters.

2. Several factors can prevent identification, including:
   a. too many paths specified in model
   b. certain kinds of model specifications can make parameters unidentified
   c. multicollinearity
   d. combination of a complex model and a small sample

3. Good news is that most software checks for identification (in something called the information matrix) and lets you know when parameters are not identified (and which ones).

USGS

13

A related, but more general problem is whether unique stable estimates for all the parameters can be obtained during estimation. The problem of identification not only applies to statistical estimation but to the evaluation of complex theoretical ideas using simple data (as discussed in

Grace, J.B., Adler, P.B., Harpole, W.S., Borer, E.T., and Seabloom, E.W. 2014 Causal networks clarify productivity–richness interrelations, bivariate plots do not. Functional Ecology, DOI: 10.1111/1365-2435 (early online) (http://onlinelibrary.wiley.com/doi/10.1111/1365-2435.12269/abstract)